ClipCap-Mamba:
Leveraging CLIP prefix to condition Mamba for image captioning

Chiara Ballanti
Sapienza University of Rome
ballanti.1844613@studenti.uniroma1.it

Abstract

Image captioning is a task that involves both computer vision and natural language processing to understand the content of an image and express it in coherent and human-readable language. In the realms of natural language processing and computer vision, transformers and attention mechanisms have been widely explored. However, this study introduces a novel perspective by diverging from the conventional focus on attention-based models. Instead, it investigates the integration of the state-of-the-art state space model *Mamba* [16] with the *ClipCap* [35] framework. This approach aims to discover novel interactions between sophisticated language models and vision-language systems, with the goal of enhancing the performance of image caption generation. The code is available on GitHub.

1 Introduction

Image captioning is a complex interdisciplinary task at the intersection of computer vision (CV) and natural language processing (NLP). It involves generating a textual description that accurately reflects the content of a given image. The generated captions are short enough to fit into a single sentence and long enough to describe all the salient contents of the image. This task demands not only the recognition of objects and actions within the image but also the ability to coherently and contextually articulate these elements using appropriate linguistic structures.

Image captioning can be conceptualized as a sequence-to-sequence learning task, where the input is an image and the output is a sentence. The primary objective is to develop models capable of bridging the gap between visual data and linguistic expression, effectively capturing the salient features of an image and translating them into a meaningful narrative.

1. A buffet table is spread with delicacies for a celebration.
2. Two women and kids standing around a table with food.
3. Four young ladies preparing food for a group.
4. Four people around a table with various plates of food.
5. A family gathered around a dinner table getting plates of food.

Figure 1: Example of an image with a possible set of captions from the *MS COCO* [29] dataset.

Image captioning models have a wide range of real-world applications, significantly enhancing the accessibility and usability of visual data across various domains. One prominent application is in assistive technology for visually impaired individuals, where image captioning systems can describe scenes, objects, and activities in real time, enabling better navigation and understanding of their surroundings. In the medical field, automated captioning of medical images facilitates disease diagnosis and treatment. Another critical application is in autonomous vehicles, where image captioning contributes to the understanding of complex driving environments by describing road scenes and traffic conditions. Additional applications can be found in service robotics, military operations, education, and image indexing.

2 Related work

In its standard configuration, the image captioning task is an image-to-sequence problem with pixels in input. These inputs are encoded as one or multiple feature vectors in the visual encoding step, which prepares the input for a second generative step, the language model. This produces a sequence of words or sub-words decoded according to a given vocabulary. Early approaches to image captioning primarily relied on encoder-decoder architectures, where a convolutional neural network (CNN) such as *AlexNet*, *VGG*, or *ResNet* was used to encode image features, and a recurrent neural network (RNN), usually a Long Short-Term Memory (LSTM) network, was employed to generate captions [52, 22, 34, 14, 42]. However, these models struggled with long-term dependencies and limited generalization to complex language structures. To address these limitations, many captioning models [57] have adopted additive attention mechanisms [5], eventually applied to image regions [4] and enhanced with spatial and semantic graphs [58, 59]. One of the most widely used techniques is self-attention, introduced by Vaswani et al. [50], which allows models to capture global dependencies in both images and text. Numerous works have explored transformer-based architectures in various directions [11, 20, 27, 38]. A few studies have applied self-attention to grid features [62] and image patches [2, 48], with Liu et al. [32] pioneering a convolution-free model using a Vision Transformer (ViT). Models like *CLIP* [41] and *SimVLM* [55] have leveraged large-scale datasets for visual encoding, while vision-and-language pre-training approaches [46, 33, 21] incorporating early fusion strategies [63] have enhanced performance. A significant advancement in multimodal processing was achieved with the development of models such as *Oscar* [26] and *VinVL* [61, 19], which integrate object tags as additional inputs to improve the alignment between vision and language. These models employ transformers to fuse visual features with textual representations, resulting in a more nuanced context understanding and improved descriptive accuracy. Concurrently, extensive research has focused on developing large-scale multimodal models, typically grounded in large language models (LLMs) and trained on vast amounts of multimodal data [7, 9, 25, 24]. In this context, image captioning has been employed as a pre-training task to enhance vision-language alignment, with further improvements in the instruction-tuning stage [31, 30]. Owing to the power of LLMs, these approaches often result in image captioning models with enhanced descriptive capabilities.

Training strategies Aligning models with human judgment is a well-known challenge in both natural language processing (NLP) and image captioning literature. To address this, various strategies for fine-tuning large language models (LLMs) have been explored. A common approach is to guide the models using a combination of input-output pairs and explicit instructions [10, 22, 54, 53]. However, LLMs frequently demonstrate a tendency to generate biased or potentially harmful content. To mitigate this issue, several works have attempted to align models with human judgment through reinforcement learning techniques [37, 45, 49, 64], while also developing methods for efficient fine-tuning that address the substantial memory demands involved in training LLMs [23, 23].

3 Datasets and benchmarks

High-quality datasets offer diverse and annotated images, enabling models to learn visual and linguistic relationships. Over time, both the quantity and quality of these datasets have grown significantly, allowing for more robust and sophisticated models. The following are examples of the most commonly used datasets in the literature:

- **COCO** (*Microsoft Common Objects in Context*) [29]. It is a large-scale dataset for object detection, segmentation, key-point detection, and captioning. The dataset consists of around 328K images of every day's complicated actions with a total of 91 item types. Each image has five captions.

- **Flickr8k** [18]. It is a collection of 8k images of people and animals gathered from the website *flick.com*. There are five captions for each of these images.

- **Flickr30k** [60]. Extension to Flickr8K, it consists of 30k images from *flick.com* with 5 captions per image. There are a total of 158K captions, and 276k manually annotated bounding boxes matching each object.

- **Nocaps** [1]. It contains 166k human-generated captions describing 15k images from the *OpenImages* validation and test sets.

- **SBU captions** [36]. It includes over 1 million images sourced from the web, each paired with a single descriptive caption. It was created by crawling image-caption pairs from various online sources, offering a vast dataset for training models to understand and describe a wide array of visual content.

- **Conceptual Captions**. It contains around 3.3 million (CC3M) [44] and 12 million (CC12M) [8] images paired with one weakly associated description automatically collected from the web with a relaxed filtering procedure.

4 Existing tools, libraries, papers with code

TensorFlow [40] and **PyTorch** [47] are popular deep learning frameworks that support building and training image captioning models. They provide modules for neural networks and efficient GPU computation. The **Hugging Face Transformers** [56] library is also widely used, offering pre-trained models and state-of-the-art architectures that can be fine-tuned for image captioning tasks. The **Natural Language Toolkit** (*NLTK*) [6] can be utilized for text processing and analysis, which is crucial for evaluating and enhancing the generated captions. Additionally, specifically designed for the evaluation of image captioning systems, the **pycocoevalcap** [43] library proves standard metrics to assess captioning performance.

5 State-of-the-art evaluation

The evaluation of state-of-the-art image captioning models involves assessing their ability to generate accurate, fluent, and contextually relevant descriptions of images. Many evaluation metrics are used to assess the performance of image captioning systems:

- **BLUE** (*Bilingual evaluation understudy*) [39]. It is a modified precision metric used to assess the quality of the machine-translated text. Individual n-grams from the produced captions are compared to a collection of reference captions, and scores are computed. The BLEU score is calculated as the number of n-grams found in both produced and reference captions.

- **METEOR** (*Metric for Evaluation of Translation with Explicit Ordering*) [13]. It is an accuracy and recall-based metric. METEOR considers the morphology of words while matching n-grams. The metric computes unigram matches at multiple levels, including exact match, porter stem match, and synonyms. After calculating all such matches, the accuracy, recall, and F-Mean are determined and used to compute the final score as $Meteor = FMean * (1 - Penalty)$.

- **ROUGE** (*Recall Oriented Understudy for Gisting Evaluation*) [28]. It is a machine translation evaluation that is based on recall. Depending on the task, several ROUGE metrics are available. *ROUGE-N*, *ROUGE-W*, *ROUGE-S*, and *ROUGE-SU* are a few of them. For image captioning, is often used **ROUGE-L**, which is based on the Longest Common Sub-sequence and evaluates the score by detecting the longest co-occurring sequence of n-grams in the sentence,

- **CIDEr** (*Consensus-based Image Description Evaluation*) [51]. It compares the resemblance of a produced caption to human-written ground truth captions. CIDEr takes into account the concepts of grammar, saliency, accuracy, and precision.

- **SPICE** (*Semantic Proportional Image Captioning Evaluation*) [3]. Because traditional n-gram assessment measures are susceptible to overlaps, SPICE is a unique image captioning metric. It analyses captions by taking the semantic information of the image into account. SPICE creates a scene graph of both the reference and produced captions to match their semantic contents.

6 Comparative evaluation

6.1 Proposed method

The *ClipCap* (CLIP Prefix for Image Captioning) [35] approach leverages the capabilities of two state-of-the-art pre-trained models: CLIP and a language model (LM). The methodology centers on utilizing the rich semantic embeddings generated by CLIP, which encapsulate key visual information, as a conditioning input for the caption

generation process. This embedding serves as a prefix to guide the generation of image captions. Then, an auto-regressive language model predicts the subsequent tokens in the caption. Since the language model cannot directly interpret CLIP embeddings, a trainable mapping network is employed to convert these embeddings into a format suitable for the LM.

CLIP The *CLIP* (Contrastive Language-Image Pretraining) model, developed by OpenAI, is a multimodal neural network designed to understand and generate both text and images by aligning them within a shared semantic space. CLIP is trained using a contrastive learning approach, where it learns to associate text descriptions with corresponding images by minimizing the distance between related text-image pairs and maximizing the distance between unrelated ones. This enables CLIP to generalize across a wide range of visual and textual tasks, including zero-shot learning on tasks for which it has not been explicitly trained.

In this setup, only the image encoder from CLIP is utilized to process visual data, extracting high-dimensional image embeddings that are then fed into a captioning model to generate descriptive text. This approach leverages the visual encoder's ability to capture rich and robust feature representations, significantly improving the accuracy and fluency of the generated captions, despite CLIP not being specifically designed for image captioning tasks.

Mamba Mamba innovatively enhances sequence modeling by incorporating a selective state space model (SSM) that achieves linear-time complexity while addressing the limitations of traditional SSMs. In a structured SSM, a one-dimensional sequence $x(t) \in \mathbb{R}^L$ is mapped to $y(t) \in \mathbb{R}^L$ through a hidden state $h(t) \in \mathbb{R}^N$, utilizing the evolution parameter $A \in \mathbb{R}^{N\times N}$ and the projection parameters $B \in \mathbb{R}^{N\times 1}$ and $C \in \mathbb{R}^{1\times N}$. The SSM is formulated using linear ordinary differential equations:

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t) \\ y(t) &= Ch(t) \end{aligned} \tag{1}$$

To integrate SSMs into deep learning frameworks, continuous-time models are discretized using a zero-order hold assumption, transforming continuous parameters A and B into their discrete counterparts $\bar{A}$ and $\bar{B}$ with a timescale parameter Δ according to

$$\begin{aligned} \bar{A} &= \exp(\Delta A) \\ \bar{B} &= (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B \end{aligned}$$

Consequently, Equation 1 can be rewritten as:

$$\begin{aligned} h'(t) &= \bar{A}h(t) + \bar{B}x(t) \\ y(t) &= Ch(t) \end{aligned} \tag{2}$$

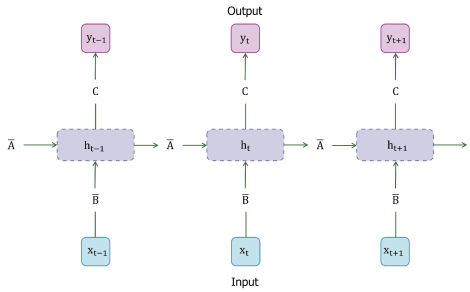


Figure 2: Illustration of a discretized SSM.

To enhance computational efficiency and scalability, Mamba synthesizes the iterative process in Equation 2 through a global convolution

$$\begin{aligned} \bar{K} &= (C\bar{B}, C\bar{A}\bar{B}, ..., \bar{A}^{L-1}\bar{B}) \\ y &= x * \bar{K} \end{aligned}$$

where L is the length of the input sequence x, $\bar{K} \in \mathbb{R}^L$ serves as the kernel of the SSM, and $*$ represents the convolution operation. To address the constraints of traditional SSMs, which exhibit fixed time-invariant parameterization, Mamba introduces a selective SSM that utilizes input-dependent matrices B, C, and timescale Δ derived from selective scans over the input sequence. This selective mechanism enhances the model's adaptability and computational efficiency by reducing redundant operations and optimizing memory usage.

The selective SSM can be implemented as a block, allowing multiple Mamba blocks to be stacked, with each block's output serving as the next block's input, as shown in Figure 3.

Compared to RNNs and LSTMs, which struggle with vanishing gradients and long-range dependencies, Mamba offers efficient computation and better memory utilization. While transformers excel in batch processing and handling long-range dependencies through attention mechanisms, they are computationally expensive, particularly during inference. Mamba addresses this by introducing

a selective SSM that incorporates input-dependent matrices to improve adaptability while preserving the computational efficiency of traditional SSMs. By bridging the gap between traditional SSMs and modern neural network architectures, Mamba delivers selective dependency handling, optimal GPU memory usage, and linear scalability with context length, making it a promising solution for a wide range of sequential data processing tasks.

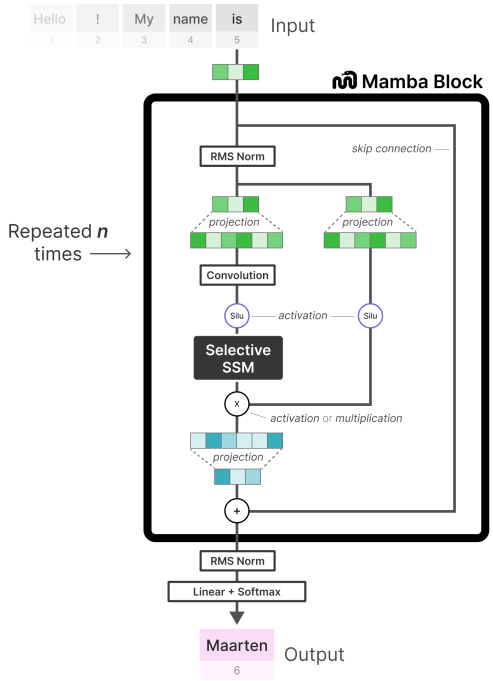


Figure 3: Overview of Mamba architecture [15].

6.2 Experiments setup

Model architecture While recent papers have proposed vision models based on the Mamba architecture, these models are still in their early stages of development and have not yet been trained on large-scale datasets. As a result, these models lack the robustness required for handling diverse visual content effectively. Therefore, to ensure high-quality visual feature extraction, the CLIP model was retained in this work. Specifically, the *CLIP-ViT-B/32* model (Vision Transformer with a patch size of 32) was used, which produces a 512-dimensional embedding for each image. CLIP, pre-trained on extensive datasets, offers strong generalization capabilities across a wide variety of images, making it a reliable choice for extracting visual features necessary for the image captioning task.

In this study, the Mamba architecture was used as both the mapping network and the language model. The mapping network consists of a fully connected layer followed by 8 stacked Mamba blocks. The fully connected layer has been introduced to transform the single embedding vector into a sequence of embeddings, aligning the input with the sequential processing requirements of the Mamba blocks. This transformation enables the Mamba architecture to effectively model and capture dependencies within the sequence, leveraging its inherent strengths as a sequence model. The output of the mapping network is then used as a prefix to condition the text generation phase of the language model. The *state-spaces/mamba-130m-hf* model from *Hugging Face* was employed for the language model.

Figure 4 shows an overview of the model architecture.

Training procedure The pre-trained CLIP model was kept frozen during training, while the mapping network and language model were fine-tuned. The final model consists of a total of 160 million trainable parameters.

The model was trained using the Flickr30k

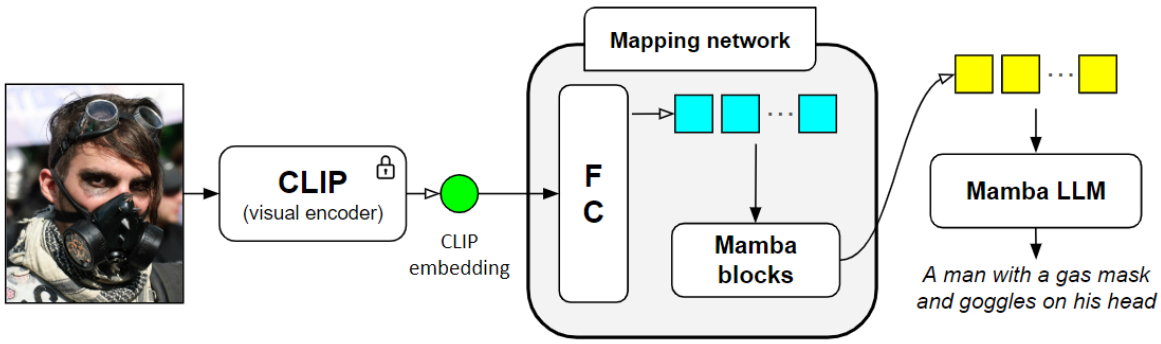


Figure 4: Overview of the proposed method architecture.

Dataset	Method	B@1	B@4	M	R	C	S
Conceptual Captions [44]	Original ClipCap [35]	-	-	-	26.71	87.26	18.5
nocaps [1]		-	-	-	65.7	11.1	-
COCO MS [29]		-	33.53	27.45	-	113.08	21.05
Flickr30k [60]	**ClipCap + Mamba**	67.6	26.9	**23.2**	**51.0**	63.8	**16.9**
	Show and Tell [52]	64.1	21.5	18.3	44.4	41.7	12.2
	Show, Attend and Tell [52]	65.6	23.6	19.2	45.4	49.1	13.3
	Up-Down [4]	**72.4**	28.3	21.6	49.5	63.3	15.9
	ORT [17]	72.2	30.1	22.8	50.4	68.8	16.9
	$\mathcal{M}^2$ Transf [11]	**72.4**	**29.8**	22.4	50.6	**68.4**	16.2

Table 1: Quantitative results. Comparison of image captioning performance of the proposed method with the original *ClipCap* and with methods trained on the *Flickr30k* caption dataset using cross-entropy loss.

dataset, which was split into training (70%), validation (20%), and test (10%) sets. A cross-entropy loss function was utilized for the optimization of the caption generation task. Model training employed the AdamW optimizer with a learning rate schedule defined by an initial warmup phase, followed by a cosine decay, with a maximum learning rate of 3×10^{-5}. Early stopping based on validation loss was employed to mitigate overfitting.

The batch size was set to 32 to allow efficient processing, and training was conducted over 50 epochs. The best model was selected based on the lowest validation loss, and its performance was subsequently assessed on the test set using all the evaluation metrics described in Section 5.

6.3 Experimental results

The quantitative and qualitative results are detailed respectively in Table 1 and in Appendix A.

Table 1 provides a comparison with several notable deep-learning-based methods trained on the Flickr30k dataset using cross-entropy loss. The results indicate that the proposed method outperforms the others in three metrics: METEOR, Rouge-L, and SPICE. These findings suggest that the model is not only generating accurate and fluent captions but also capturing meaningful semantic content. Unfortunately, none of the recent and promising image captioning methods use the Flickr30k dataset, making direct comparisons impossible. Table 1 also includes results from the original Clipcap for reference, although a direct comparison is not feasible due to dataset differences. Nevertheless, the results are consistent with those reported in the original Clipcap study.

6.4 Discussion

The obtained results are promising. The model can understand and generate coherent captions describing the content of the images.

In most cases, the model accurately identifies the primary objects in the pictures. However, in more complex images, it tends to focus solely on the main scene or action, often overlooking background details or surrounding events. This highlights the strength of Clip in object recognition but also reveals the limitations of the small 512-dimensional embedding, which must compress a significant amount of information.

For generating the final caption, the model can use both greedy and beam search techniques. In only a few cases does the model tend to hallucinate, and this mostly occurs with actions rather than objects. Interestingly, this issue typically appears in just one of the two generated captions. Therefore, generating both captions can help mitigate this problem.

Illumination conditions are a critical factor that

can influence the performance and reliability of the generated captions. However, poor or uneven lighting does not significantly impact the model's ability to detect objects. Even the presence of shadows or reflections, which could cause objects to blend into the background, rarely poses a problem. For instance, in a sunset scene where objects and people become black silhouettes, or with a large metallic object that reflects and duplicates the entire scene, the model-generated captions remain unaffected.

The model faces two main challenges. It struggles with accurately identifying a person's gender or age, particularly when multiple people are present in the scene. Although the correct number of people is often detected, at least one individual is misgendered. For example, men might be labeled as women, and vice versa, or children might be mistaken for adults. Another challenge involves occasional confusion in matching colors with respective objects, especially when objects are too close to each other, leading to color inversions.

7 Conclusions and future work

In conclusion, this study successfully integrates the state-of-the-art state space model, Mamba, with the ClipCap framework, demonstrating the potential of non-attention-based models in image captioning. The results suggest that Mamba can effectively compete with traditional transformer-based architectures, producing coherent and human-readable image captions with comparable performance.

Future work could build on these findings in several ways. First, training the model for more epochs may further improve its performance. Additionally, training on different datasets could help validate the model's generalization across various domains. A crucial step for future research would involve training the model on the same dataset used in the original ClipCap framework, enabling a direct comparison and providing clearer insights into the effectiveness of the Mamba-based approach against traditional attention-based models. Due to current GPU limitations, this study was constrained to using CLIP ViT-B/32 and Mamba-130M. Future experiments with more complex versions of CLIP and larger pre-trained Mamba models could yield more robust results. Additionally, the recently released Mamba2 [12] could be tested. Further research might also explore hybrid models that combine the strengths of both attention-based and state space models, potentially advancing image captioning performance even further.

Overall, this study highlights the viability of state space models as a competitive alternative to transformers in complex tasks such as image captioning.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 3, 6

[2] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 2

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 3

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2, 6

[5] Dzmitry Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 2

[6] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 3

[7] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wikillava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1818–1826, 2024. 2

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 3

[9] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 2

[11] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020. 2, 6

[12] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024. 7

[13] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 3

[14] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[15] Maarten Grootendorst. A visual guide to mamba and state space models. *Exploring Language Models*, 19 Feb 2024. 5

[16] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1

[17] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in neural information processing systems*, 32, 2019. 6

[18] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 2

[19] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17980–17989, 2022. 2

[20] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643, 2019. 2

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[22] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[23] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 2

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *In-*

ternational conference on machine learning, pages 19730–19742. PMLR, 2023. 2

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[26] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 2

[27] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17990–17999, 2022. 2

[28] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 3

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2, 6

[30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[32] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*, 2021. 2

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[34] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014. 2

[35] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 1, 3, 6

[36] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 3

[37] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 2

[38] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10971–10980, 2020. 2

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 3

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 3

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[42] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 2

[43] salaniz. pycocoevalcap: Microsoft coco caption evaluation. https://github.com/salaniz/pycocoevalcap, 2020. GitHub repository. 3

[44] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 3, 6

[45] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[46] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[47] TensorFlow Team. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015. *URL http://tensorflow. org/. Software available from tensorflow. org*, 2015. 3

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2

[49] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[50] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[51] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 3

[52] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2, 6

[53] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 2

[54] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*, 2022. 2

[55] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[56] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020. 3

[57] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan

Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[58] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019. 2

[59] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018. 2

[60] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2, 6

[61] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 2

[62] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15465–15474, 2021. 2

[63] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049, 2020. 2

[64] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. 2

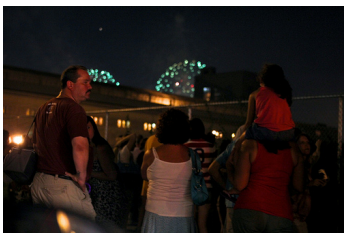
A Qualitative results

A.1 Successes

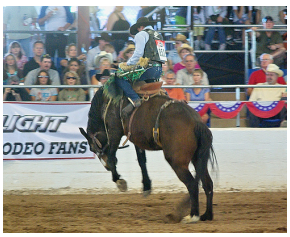
(a) A man is playing soccer at sunset.

(b) A man is standing in front of a building with a lot of garbage on the ground.

(c) A man in a white shirt is working on a large metal structure.



(d) A group of people are watching fireworks at night.



(e) A man with a cowboy hat is riding a bucking bronco at a rodeo.

(f) Two dogs wearing pink leotards are working down the street.

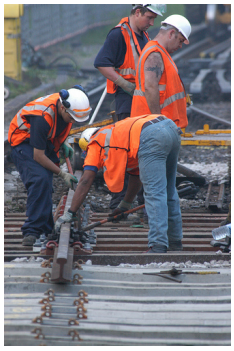
(g) A group of children are walking in a line.

(h) A group of people are riding on the back of a truck.

(i) A group of people are sitting at a table in front of a restaurant.

(j) A young boy in a blue shirt and jeans is running on a track.

(k) A man is walking down the street in front of a monument.



(l) Workers in orange vests and hard hats are working on a railroad track.

(m) A woman is sewing a garment.

Figure 5: Examples of successfully captioned images.

A.2 Failures

(a) A girl in a blue tank top is sitting on a bench.

(b) A woman in a red shirt is standing in front of a row of red and white laundromats.

(c) A man is petting a cat.

(d) Two boys are hugging each other on the road.

(e) A woman and a man are walking down the street.

(f) A man in a black shirt and a woman in a black shirt are playing instruments.

(g) Two people are riding on a roller coaster.

(h) A man in a yellow shirt and a man in a red shirt are running in a race.

(i) A group of men are playing volleyball on a beach.

(j) A man in a blue shirt is working on a pottery wheel.

(k) A woman and two children are standing in a room.

(l) A man in a blue uniform is working on a piece of equipment.

(m) A little girl in a green shirt is standing in front of a mirror.

Figure 6: Examples of incorrectly captioned images.